Medical Image Analysis 36 (2017) 123–134

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data

Mingxia Liu[a], Jun Zhang[a], Pew-Thian Yap[a], Dinggang Shen[a,b,*]

[a] Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, NC, USA
[b] Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

ARTICLE INFO

Article history:
Received 11 April 2016
Revised 31 October 2016
Accepted 7 November 2016
Available online 16 November 2016

Keywords:
Multi-modality
Incomplete data
Alzheimer's disease
Classification

ABSTRACT

Effectively utilizing incomplete multi-modality data for the diagnosis of Alzheimer's disease (AD) and its prodrome (*i.e.*, mild cognitive impairment, MCI) remains an active area of research. Several multi-view learning methods have been recently developed for AD/MCI diagnosis by using incomplete multi-modality data, with each view corresponding to a specific modality or a combination of several modalities. However, existing methods usually ignore the underlying coherence among views, which may lead to sub-optimal learning performance. In this paper, we propose a view-aligned hypergraph learning (VAHL) method to explicitly model the coherence among views. Specifically, we first divide the original data into several views based on the availability of different modalities and then construct a hypergraph in each view space based on sparse representation. A view-aligned hypergraph classification (VAHC) model is then proposed, by using a view-aligned regularizer to capture coherence among views. We further assemble the class probability scores generated from VAHC, via a multi-view label fusion method for making a final classification decision. We evaluate our method on the baseline ADNI-1 database with 807 subjects and three modalities (*i.e.*, MRI, PET, and CSF). Experimental results demonstrate that our method outperforms state-of-the-art methods that use incomplete multi-modality data for AD/MCI diagnosis.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease, characterized by progressive impairment of neurons and synaptic functioning. As an increasingly prevalent disease, AD is regarded as a major world-wide challenge to global health care systems (Brookmeyer et al., 2007). The total estimated prevalence of AD is expected to be 13.8 million in the United States by 2050 (Association et al., 2013). It is reported that the direct cost of care for AD patients provided by family members and health-care systems is more than $100 billion per year (Association et al., 2013). In recent years, much effort has been made to find early diagnostic markers to evaluate AD risk pre-symptomatically in a rapid and rigorous way, allowing early interventions that may prevent or at least delay the onset of AD, as well as its prodrome, *i.e.*, mild cognitive impairment (MCI) (Reiman et al., 2010).

Recent research and clinical studies have shown that structural magnetic resonance imaging (MRI), fluorodeoxyglucose positron emission tomography (FDG-PET) and cerebrospinal fluid (CSF) are among the best-established data modalities to identify biomarkers for AD progression and pathology (Reiman et al., 2010). Specifically, structural MRI provides anatomical information about the brain, and feature representations generated from MRI (*e.g.*, cortical thickness, regional volumetric measures, and connectivity information) can be used to quantify AD-associated brain abnormalities (Jack et al., 2008; Cuingnet et al., 2011; Wolz et al., 2011; Liu et al., 2016; Zhang et al., 2016). Also, FDG-PET (PET for short) can be employed to detect the abnormality in cerebral metabolic rate for glucose in human brain (Chetelat et al., 2003; Herholz et al., 2002; Foster et al., 2007). In addition, CSF total-tau (t-tau), CSF tau hyperphosphorylated at threonine 181 (p-tau) and the decrease of CSF amyloid β (Aβ) are closely related to the cognitive decline in AD and MCI subjects (Hansson et al., 2006; Kawarabayashi et al., 2001). In the literature, extensive studies have shown that multi-modality data (*e.g.*, MRI, PET and CSF) provide complementary information that can improve the performance of AD/MCI diagnosis (Ingalhalikar et al., 2012; Yuan et al., 2012; Xiang et al., 2014; Thung et al., 2014). However, the problem of incomplete data remains a big challenge in making use of multi-modality data, since there may be missing values existing in some modalities due to poor data quality and patient dropouts. For instance, while baseline MRI data are fully available for all subjects in the Alzheimer's

* Corresponding author.
E-mail addresses: mxliu1226@gmail.com (M. Liu), dgshen@med.unc.edu (D. Shen).

http://dx.doi.org/10.1016/j.media.2016.11.002
1361-8415/

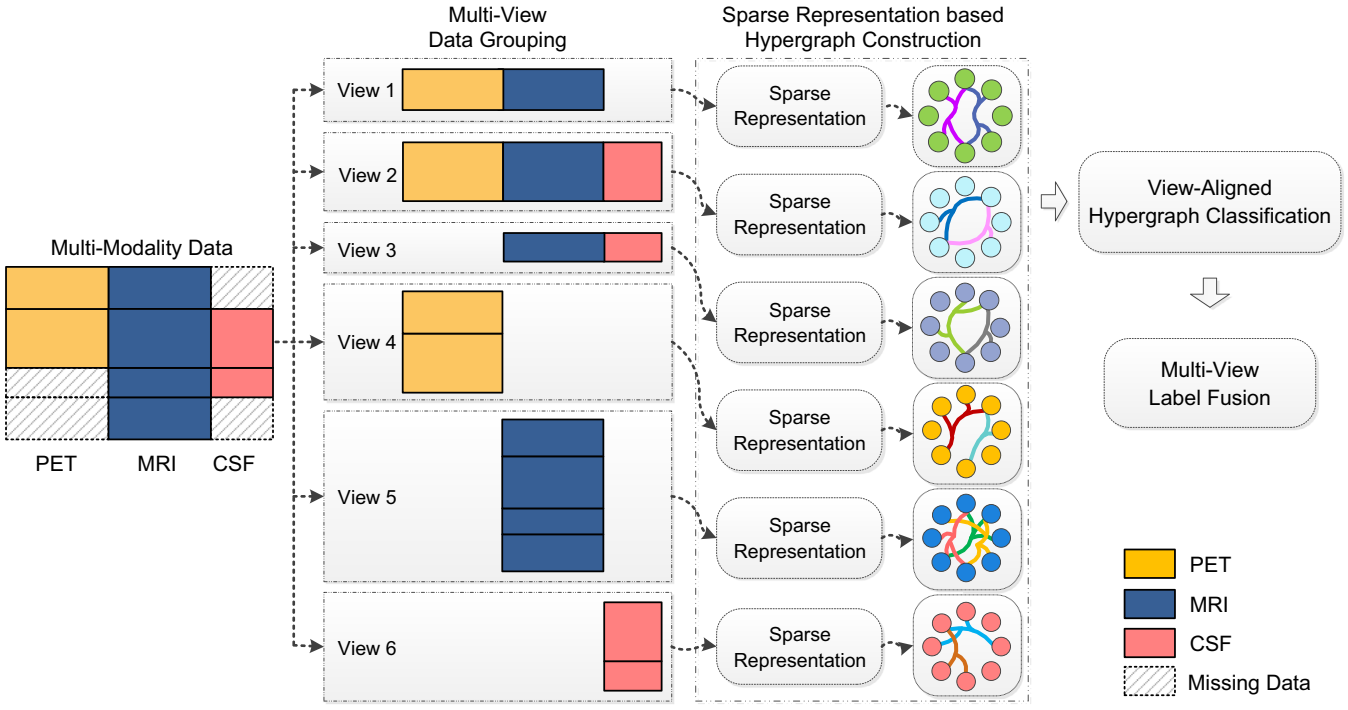

Fig. 1. Illustration of the proposed view-aligned hypergraph learning method, where subjects from the baseline ADNI-1 database are taken as examples. Subjects are divided into M ($M = 6$ in this study) views according to the data availability of a certain combination of modalities, where each view contains subjects with complete data of combined modalities. We then compute the distances among subjects via a sparse representation model, and construct one hypergraph in each view space. A view-aligned hypergraph classification method is further proposed, followed by a multi-view label fusion method to make a final classification decision.

Disease Neuroimaging Initiative (ADNI) database (Jack et al., 2008), PET and CSF data are only available for roughly half the subjects.

Currently, several approaches have been developed to handle incomplete multi-modality data (Hastie et al., 1999; Schneider, 2001; Golub and Reinsch, 1970; Yuan et al., 2012; Xiang et al., 2014; Thung et al., 2014). In general, existing methods can be divided into three categories, *i.e.*, sample exclusion methods, imputation methods, and multi-view methods. Sample exclusion methods discard subjects with incomplete data from the study, leading to sub-optimal performance due to potentially insufficient sample size (Hastie et al., 2005). Imputation methods estimate missing values based on available data using specific imputation techniques, *e.g.*, expectation maximization (EM) (Schneider, 2001), singular value decomposition (SVD) (Golub and Reinsch, 1970), and matrix completion (Thung et al., 2014). However, the effectiveness of these approaches can be affected by imputation artifacts. Without discarding subjects or imputing missing values, several recently developed multi-view learning methods (Yuan et al., 2012; Xiang et al., 2014) demonstrate greater accuracies in AD/MCI diagnosis. Multi-view methods generally divide the data into several views, with each view corresponding to a modality or a combination of modalities. Diagnosis is then performed using a multi-view learning algorithm. However, these approaches usually ignore the underneath coherence among views. Integrating these views coherently is expected to achieve better diagnostic performance.

In this paper, we propose a view-aligned hypergraph learning (VAHL) method that utilizes incomplete multi-modality data for AD/MCI diagnosis. Compared with conventional methods, VAHL explicitly incorporates the coherence among views into the learning model, where the optimal weight for each view can also be learned from the data automatically. Fig. 1 presents a schematic diagram of the proposed framework using subjects in ADNI-1 database with block-wise missing features (Xiang et al., 2014; Yuan et al., 2012). We first divide the whole data set into M views ($M = 6$ in Fig. 1) consisting of combinations of modalities. We compute the distances among subjects using a sparse representation model and then construct one hypergraph in each view space. We further propose a view-aligned hypergraph classification model, where the coherence among views is explicitly captured by a proposed view-aligned regularizer. The basic idea of such view-aligned regularizer is that, for one subject represented by two feature vectors in two view spaces, the estimated class labels for such two feature vectors should be similar because they denote the same subject. To arrive at a final classification decision, we agglomerate the class probability scores obtained from different views, via a multi-view label fusion method.

The rest of the paper is organized as follows. In Section 2, we describe the data used in this study and flesh out the proposed method. In Section 3, we describe the methods used for comparison, the experimental settings, and the experimental results based on the baseline ADNI database (Jack et al., 2008). In Section 4, we investigate the learned weights for different views, the influence of parameters on the classification performance, as well as the influence of the proposed sparse representation based distance measurement for constructing hypergraphs. In Section 5, we conclude this paper and discuss possible future research directions.

2. Material and method

In this section, we first introduce the database and image pre-processing pipeline used in this study (Section 2.1), and then present the proposed view-aligned hypergraph learning (VAHL) method, which includes multi-view data grouping (Section 2.2), sparse representation based hypergraph construction (Section 2.3), view-aligned hypergraph classification (Section 2.4), and multi-view label fusion (Section 2.5).

Table 1
Demographic and clinical information of subjects in the baseline ADNI-1 database.

	AD	MCI	NC
Male/Female	99/87	254/141	118/108
Age (Mean ± SD)	75.40 ± 7.60	74.90 ± 7.30	76.00 ± 5.00
Edu. (years) (Mean ± SD)	14.70 ± 3.10	15.70 ± 3.00	16.00 ± 2.90
MMSE (Mean ± SD)	23.30 ± 2.00	27.00 ± 1.80	29.10 ± 1.00
CDR (Mean ± SD)	0.75 ± 0.25	0.50 ± 0.03	0.00 ± 0.00

Note: Values reported as Mean ± Stand Deviation (SD); MMSE: mini-mental state examination; CDR: Clinical Dementia Rating.

2.1. Subjects and data pre-processing

The ADNI-1 database (Jack et al., 2008) is used in this study. According to the Mini-Mental State Examination (MMSE) scores, subjects in ADNI-1 can be divided into three categories: normal control (NC) subjects, MCI subjects, and AD subjects. The general inclusion/exclusion criteria used by ADNI-1 are summarized as follows: 1) NC subjects: Mini-Mental State Examination (MMSE) scores between 24 and 30 (inclusive), a Clinical Dementia Rating (CDR) of 0, non-depressed, non MCI and non-demented; 2) MCI subjects: MMSE scores between 24 and 30 (inclusive), a memory complaint, have objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living and an absence of dementia; 3) mild AD: MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0 and meets NINCDS/ADRDA criteria for probable AD. In addition, some MCI subjects had converted to AD within 24 months, while some other MCI subjects were stable over time. According to whether MCI subjects would convert to AD within 24 months, the MCI subjects are divided into two categories: 1) stable MCI (sMCI) subjects, if diagnosis was MCI at all available time points (0–96 months); 2) progressive MCI (pMCI) subjects, if diagnosis was MCI at baseline but these subjects converted to AD after baseline within 24 months.

In the baseline ADNI-1 database, there are a total of 807 subjects, including 186 AD subjects, 226 NCs and 395 MCI subjects (consisting of 169 pMCI subjects and 226 sMCI subjects). Detailed description for each category can be found at website.[1] It is worth noting that all subjects in the baseline ADNI-1 database have T1-weighted structural MRI data, while only 396 subjects have FDG-PET data and 406 subjects have CSF data. The demographic information of the studied subjects (*i.e.*, gender, age, and education) and clinical scores (*i.e.*, MMSE and CDR global) used in this study are summarized in Table 1.

We extract features based on regions-of-interest (ROIs) from MR and PET images. Specifically, for each MR image, we apply the anterior commissure (AC)-posterior commissure (PC) correction using the MIPAV software package.[2] We then re-sample the images to 256 × 256 × 256 resolution, and apply the N3 algorithm (Sled et al., 1998) to correct intensity inhomogeneity. Skull stripping (Wang et al., 2011) is then performed, followed by manual editing to ensure that both skull and dura are cleanly removed. Next, we remove the cerebellum by warping a labeled template to each skull-stripped image. Afterwards, FAST (Zhang et al., 2001) in the FSL software package [3] is then applied to segment the human brain into three different tissue types, *i.e.*, gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF). Meanwhile, the anatomical automatic labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002), with 90 pre-defined ROIs in the cerebrum, are aligned to the native space of each subject using a deformable registration algorithm, *i.e.*, HAMMER (Shen and Davatzikos, 2002) that is also extended and applied to other applications (Qiao et al., 2009; Yang et al., 2008; Xue et al., 2006; Verma et al., 2005). Finally, for each subject, we extract the volumes of GM tissue inside those 90 ROIs as features, normalized by the total intracranial volume (estimated by the summation of GM, WM, and CSF volumes from all ROIs). For PET images, we first align each PET image onto its corresponding MR image via a rigid registration, and then compute the mean intensity of each ROI in the PET image as features. In this study, we also employ five CSF biomarkers, including amyloid β (Aβ42), CSF total tau (t-tau), CSF tau hyperphosphorylated at threonine 181 (p-tau), and two tau ratios with respect to Aβ42 (*i.e.*, t-tau/Aβ42 and p-tau/Aβ42). Ultimately, we have a 185-dimensional feature vector for a subject with complete data, including 90 MRI features, 90 PET features and 5 CSF features.

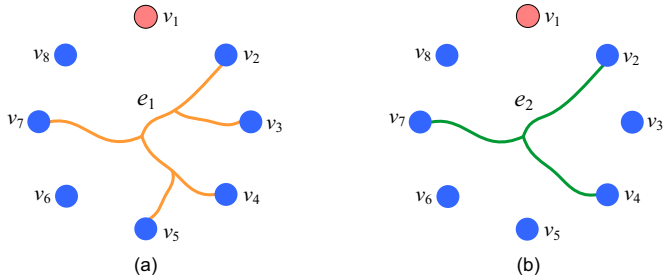

Fig. 2. Illustration of the proposed hyperedge construction method. Two hyperedges (*i.e.*, e_1 and e_2) are built by connecting the centroid vertex v_1 with the other vertices, according to the sparse representation coefficients obtained by using two l_1 regularization parameters.

2.2. Multi-view data grouping

For subjects with block-wise incomplete MRI, PET and CSF data in the baseline ADNI-1 database, we group them into $M(M = 6)$ views, including "PET+MRI", "PET+MRI+CSF", "MRI+CSF", "PET", "MRI", and "CSF". As shown in Fig. 1, subjects in View 1 have both PET and MRI features, while those in View 6 only have CSF data. In this way, we have complete feature representations for each subject in each view. Using such data grouping strategy, we can make full use of all subjects, without discarding any subjects with missing data or imputing those missing values. Such data grouping method is also used in Yuan et al. (2012) and Xiang et al. (2014) for problems with block-wise incomplete multi-modality data.

The purpose of such multi-view data grouping strategy is to fully utilize all subjects, by grouping them into different views according to the availability of data modalities. Currently, this data grouping approach can only be applied to block-wise incomplete data problem. For more general problems where there may be some missing values in a specific modality for some subjects, we can first impute these missing values using some simple technique (*e.g.*, EM or SVD), and then group subjects into different views.

2.3. Sparse representation based hypergraph construction

In this study, AD/MCI diagnosis is formulated as a hypergraph based multi-view learning problem. A hypergraph is a generalization of the traditional graph, where each edge (called hyperedge) is a non-empty subset of the vertex set (Zhou et al., 2006; Gao et al., 2012). As shown in Fig. 2 (a), the hyperedge e_1 contains 5 vertices (*i.e.*, v_2, v_3, v_4, v_5, and v_7), which demonstrates some high-order relationship among vertices. In contrast, an edge in a conventional graph can only convey the pairwise relationship by connecting only two vertices. For the convenience of presentation, we now

[1] http://adni.loni.usc.edu.
[2] http://mipav.cit.nih.gov/index.php.
[3] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki.

introduce some notations for hypergraphs. Throughout the paper, we denote matrices, vectors, and scalars using boldface upper-case letters, boldface lower-case letters, and normal italic letters, respectively. Let $\mathcal{G}^m = (\mathcal{V}, \mathcal{E}^m, \mathbf{w}^m)$ denote the mth hypergraph corresponding to the mth ($m = 1, 2, \ldots, M$) view, where $\mathcal{V}$ represents the vertex set that contains N vertices, $\mathcal{E}^m$ denotes the hyperedge set, and $\mathbf{w}^m$ is the weights for hyperedges (with the element $w^m_{e_j}$ representing the weight for the hyperedge e_j in the mth view space). We let N^m_e represent the number of hyperedges in the mth hypergraph. Denote $\mathbf{W}^m \in \mathbb{R}^{N^m_e \times N^m_e}$ as a diagonal matrix of hyperedge weights, *i.e.*, $W^m_{j,j} = w^m_{e_j}$. That is, each diagonal element of $\mathbf{W}^m$ denotes the weight for a specific hyperedge in the classification task, with a larger value representing that the hyperedge is more important. Let $\mathbf{H}^m \in \mathbb{R}^{N \times N^m_e}$ denote the vertex-hyperedge incidence matrix, with the (v_n, e_j)-entry indicating whether the vertex v_n is connected with other vertices in the hyperedge e_j, *e.g.*,

$$h^m_{v_n,e_j} = \begin{cases} 1, & \text{if } v_n \in e_j, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The degree of a vertex v_n is defined as

$$d^m_{v_n} = \sum_{e_j \in \mathcal{E}^m} w^m_{e_j} h^m_{v_n,e_j}, \tag{2}$$

and the degree for a hyperedge e_j is defined as

$$\delta^m_{e_j} = \sum_{v_n \in \mathcal{V}} h^m_{v_n,e_j}. \tag{3}$$

A key point for hypergraph learning is constructing a set of hyperedges to efficiently model the structure information of data. In conventional methods, the Euclidean distance is generally used to indicate the similarity between pairs of vertices for constructing hyperedges. For instance, in the star expansion method (Zien et al., 1999), we first select each vertex as the centroid vertex, and then construct a hyperedge by connecting this centroid vertex to its s nearest neighbor vertices, where the similarity between two vertices is evaluated by the Euclidean distance. However, the Euclidean distance can only model the local structure information among vertices and does not utilize global information. To address this problem, we propose a sparse representation based distance measurement for hyperedge construction. The reason we utilize sparse representation for computing similarities among vertices is that sparse representation coefficients have proven to be effective in reflecting the global data structure and also robust to data noise (Wright et al., 2009; Qiao et al., 2010).

Given a set of training samples $\{\mathbf{x}_n\}_{n=1}^N$ with $\mathbf{x}_n \in \mathbb{R}^D$, the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ contains N samples in its columns. The goal of sparse representation (Qiao et al., 2010) is to represent each $\mathbf{x}_n$ using as few samples as possible. Hence, we expect to seek a sparse representation weight vector $\mathbf{s}_n$ for each $\mathbf{x}_n$ via the following modified l_1 minimization problem

$$\begin{aligned} &\min_{\mathbf{s}_n} \|\mathbf{x}_n - \mathbf{X}\mathbf{s}_n\| + \beta \|\mathbf{s}_n\|_1 \\ &\text{s.t. } 1 = \mathbf{1}^\top \mathbf{s}_n, \end{aligned} \tag{4}$$

where $\mathbf{s}_n = [s_{n,1}, \ldots, s_{n,n-1}, 0, s_{n,n+1}, \ldots, s_{n,N}]^\top$ is an N-dimensional vector where the nth element is equal to zero (implying that $\mathbf{x}_n$ is removed from $\mathbf{X}$). Note that the element $s_{n,j}$ ($j \neq n$) denotes the contribution of $\mathbf{x}_j$ to the reconstruction of $\mathbf{x}_n$. The regularization parameter β is used to control the sparsity of $\mathbf{s}_n$, and $\mathbf{1} \in \mathbf{R}^N$ is a vector of all ones. In Eq. (4), the weight vector $\hat{\mathbf{s}}_n$ is computed globally in terms of samples from all classes, naturally characterizing the importance of the other samples for the reconstruction of $\mathbf{x}_n$. In other words, sample $\mathbf{x}_n$ is mainly associated with only a few samples with prominent non-zero coefficients in its reconstruction.

With the optimal weight vector $\hat{\mathbf{s}}_n$ for each $\mathbf{x}_n$ ($n = 1, 2, \ldots, N$) learned from Eq. (4), the sparse representation weight matrix $\mathbf{S}$ is defined as

$$\mathbf{S} = [\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_n, \ldots, \hat{\mathbf{s}}_N]^\top. \tag{5}$$

Based on the sparse representation coefficients in Eq. (5), we adopt the star expansion algorithm (Zien et al., 1999) to generate a set of hyperedges. Specifically, in each view space, we first select each vertex as the centroid vertex, and then construct a hyperedge by connecting this centroid vertex to the other vertices, with the sparse representation coefficients as similarity measure. That is, a large coefficient demonstrates a strong connectivity, and a zero coefficient denotes no connectivity. The element $h^m_{v_n,e_j}$ of the vertex-hyperedge incidence matrix $\mathbf{H}^m$ is defined as

$$h^m_{v_n,e_j} = \begin{cases} |S_{n,j}|, & \text{if } |S_{n,j}| > \theta, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where θ is a small threshold (which is set to 0.001 empirically in this study), and $S_{n,j}$ is the (n, j)-entry of $\mathbf{S}$ in Eq. (5).

It is worth noting that a larger β in Eq. (4) will lead to more zeros in the representation coefficients, which indicates that fewer vertices are used to represent the centroid vertex. In this way, the corresponding hyperedge would contain less vertices, demonstrating a relatively local data structure. To model multi-scale structure information of data, we propose to employ multiple (*e.g.*, q) values for β to construct multiple sets of hyperedges. As illustrated in Fig. 2, we construct two hyperedges (*i.e.*, e_1 and e_2) by connecting a centroid vertex v_1 with the other vertices, where each hyperedge corresponds to a specific β. For the hypergraph $\mathcal{G}^m$ in the mth view space, we can finally obtain $N^m_e = qN$ hyperedges in the vertex-hyperedge incidence matrix $\mathbf{H}^m$. In this way, we can obtain hundreds of hyperedges, some of which may not be informative enough for subsequent classification model. We further propose to learn optimal weights for hyperedges in Section 2.4.2 in order to identify those most informative hyperedges.

2.4. View-aligned hypergraph classification

In the following, we first propose a view-aligned regularizer to explicitly model the underlying coherence among views, and then develop a view-aligned hypergraph classification model as well as an efficient alternating optimization algorithm.

2.4.1. View-aligned regularizer

Denote $\mathbf{f}^m \in \mathbb{R}^N$ as the class probability score vector for N subjects in the mth view, and $\mathbf{F} = [\mathbf{f}^1, \mathbf{f}^2, \cdots, \mathbf{f}^m, \ldots, \mathbf{f}^M] \in \mathbb{R}^{N \times M}$, where M is the number of views. The proposed view-aligned regularizer is illustrated in Fig. 3, where different colors and shapes denote different views and subjects, respectively. For instance, circles represent a subject having PET (View 4), MRI (View 5) and CSF (View 6) data, that are denoted as $\mathbf{x}^4_1$, $\mathbf{x}^5_1$ and $\mathbf{x}^6_1$, respectively. Intuitively, after being mapped into the label space, their estimated class probability scores (*i.e.*, f^4_1, f^5_1, and f^6_1) should be close to each other, since they represent the same subject. Similarly, for the subject with only PET and MRI features (*i.e.*, triangles for $\mathbf{x}^4_2$ and $\mathbf{x}^5_2$), the distance between f^4_2 and f^5_2 should be small in the label space. Denote $\mathbf{\Omega}^m \in \mathbb{R}^{N \times N}$ as a diagonal matrix, with the diagonal element $\Omega^m_{n,n} = 0$ if the nth subject has missing values in the mth view, and $\Omega^m_{n,n} = 1$, otherwise. Then, the proposed view-aligned regularizer is defined as

$$\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{p=1}^{M} \Omega^m_{n,n}\Omega^p_{n,n}(f^m_n - f^p_n)^2 = \sum_{m=1}^{M} (\mathbf{f}^m)^\top \mathbf{\Omega}^m \sum_{p=1}^{M} \mathbf{\Omega}^p (\mathbf{f}^m - \mathbf{f}^p). \tag{7}$$

Let $\mathbf{D}^m_v$ represent the vertex degree matrix whose diagonal entries correspond to the degree of each vertex. Denote $\mathbf{D}^m_e$ as

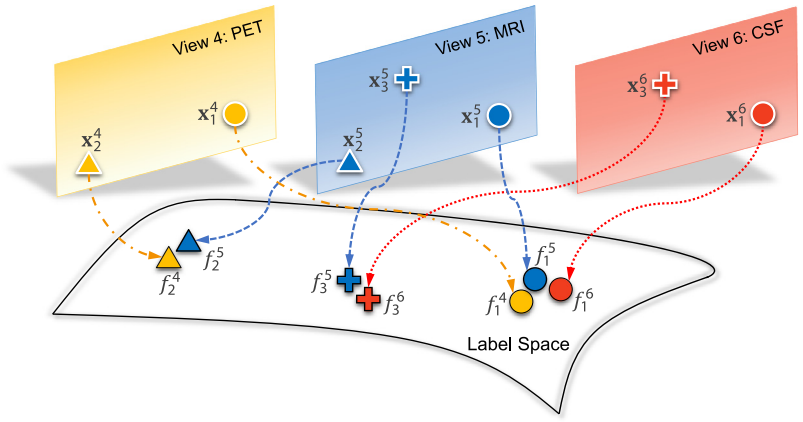

Fig. 3. Illustration of the proposed view-aligned regularizer. Circle, cross and triangle represent three subjects, respectively. Yellow, blue, and red denote the views of "PET", "MRI", and "CSF", respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the hyperedge degree matrix, with diagonal elements representing the degree of each hyperedge. The hypergraph regularization term (Zhou et al., 2006) is defined as

$$\frac{1}{2}\sum_{m=1}^{M}\sum_{e_j\in\mathcal{E}^m}\sum_{v_n,v_i\in\mathcal{V}}\frac{w^m_{e_j}h^m_{v_n,e_j}h^m_{v_i,e_j}}{\delta^m_{e_j}}\times\left(\frac{f^m_{v_n}}{\sqrt{d^m_{v_n}}}-\frac{f^m_{v_i}}{\sqrt{d^m_{v_i}}}\right)^2 = \sum_{m=1}^{M}(\mathbf{f}^m)^\top\mathbf{L}^m\mathbf{f}^m, \tag{8}$$

where $\mathbf{L}^m=\mathbf{I}-\mathbf{\Theta}^m$ is the hypergraph Laplacian matrix, $\mathbf{I}$ is an identity matrix, and $\mathbf{\Theta}^m=(\mathbf{D}^m_v)^{-\frac{1}{2}}\mathbf{H}^m\mathbf{W}^m(\mathbf{D}^m_e)^{-1}(\mathbf{H}^m)^\top(\mathbf{D}^m_v)^{-\frac{1}{2}}$.

2.4.2. View-aligned hypergraph classification

Denote $\mathbf{y}=[(\mathbf{y}^{la})^\top,(\mathbf{y}^{un})^\top]^\top\in\mathbb{R}^N$, where $\mathbf{y}^{la}$ denotes label information for labeled data and $\mathbf{y}^{un}$ represents label information for unlabeled data. For the nth sample, $y_n=1$ if it is associated with the positive class (*e.g.*, AD), $y_n=-1$ if it belongs to the negative class (*e.g.*, NC), and $y_n=0$ if its category is unknown. Since different views and hyperedges may play different roles in a classification task, it is intuitively reasonable to learn weights for different views and for hyperedges from data. Denote $\boldsymbol{\alpha}\in\mathbb{R}^M$ as a weight vector, with its element α^m representing the weight for the mth view. Denote the Frobenius norm of the matrix $\mathbf{W}^m$ as $\|\mathbf{W}^m\|_F^2=\sum_{i,j}|W^m_{i,j}|^2$. In this study, we resort to the multi-task learning framework (Argyriou et al., 2008) for classification, and regard the classification in each view space as a specific learning task. The proposed view-aligned hypergraph classification (VAHC) model is formulated as

$$\begin{aligned}\min_{\mathbf{F},\boldsymbol{\alpha},\{\mathbf{W}^m\}_{m=1}^M}\ & \sum_{m=1}^{M}\|\mathbf{\Omega}^m(\mathbf{f}^m-\mathbf{y})\|^2+\sum_{m=1}^{M}(\alpha^m)^2(\mathbf{f}^m)^\top\mathbf{L}^m\mathbf{f}^m\\ &+\mu\sum_{m=1}^{M}(\mathbf{f}^m)^\top\mathbf{\Omega}^m\sum_{p=1}^{M}\mathbf{\Omega}^p(\mathbf{f}^m-\mathbf{f}^p)+\lambda\sum_{m=1}^{M}\|\mathbf{W}^m\|_F^2,\\ \text{s.t.}\ & \sum_{m=1}^{M}\alpha^m=1,\ \forall\,\alpha^m\geq 0;\\ & \sum_{j=1}^{N^m_e}W^m_{j,j}=1,\ \forall\,W^m_{j,j}\geq 0,\end{aligned} \tag{9}$$

where the first term is the empirical loss, and the second one is the hypergraph Laplacian regularizer (Zhou et al., 2006). It is worth noting that the third term in Eq. (9) is the proposed view-aligned regularizer, encouraging the similarity of the estimated class labels for one subject represented in two different views. The last term and those constraints in Eq. (9) are used to penalize the complexity of the weights (*i.e.*, $\mathbf{W}^m$) for hyperedges and also the weights (*i.e.*, $\boldsymbol{\alpha}$) for views. The regularization parameter $(\alpha^m)^2$ is used to prevent the degenerate solution of $\boldsymbol{\alpha}$. In addition, μ and λ are regularization parameters for our proposed view-aligned regularizer and the hyperedge weight regularizer, respectively. With Eq. (9), one can jointly learn the class probability scores $\mathbf{F}$, the optimal weights for different views (*i.e.*, $\boldsymbol{\alpha}$), and the optimal weights for hyperedges (*i.e.*, $\{\mathbf{W}^m\}_{m=1}^M$) from data.

Since the problem in Eq. (9) is not jointly convex with respect to $\mathbf{F}$, $\boldsymbol{\alpha}$, and $\{\mathbf{W}^m\}_{m=1}^M$, we adopt an alternating optimization method to solve the proposed objective function. Specifically, in the first step, we aim to optimize $\mathbf{F}$ with fixed $\boldsymbol{\alpha}$ and $\{\mathbf{W}^m\}_{m=1}^M$. In such case, the objective function in Eq. (9) can be written as

$$\begin{aligned}\min_{\mathbf{F}}\ & \sum_{m=1}^{M}\|\mathbf{\Omega}^m(\mathbf{f}^m-\mathbf{y})\|^2+\sum_{m=1}^{M}(\alpha^m)^2(\mathbf{f}^m)^\top\mathbf{L}^m\mathbf{f}^m\\ &+\mu\sum_{m=1}^{M}(\mathbf{f}^m)^\top\mathbf{\Omega}^m\sum_{p=1}^{M}\mathbf{\Omega}^p(\mathbf{f}^m-\mathbf{f}^p).\end{aligned} \tag{10}$$

The partial derivative of the objective function in Eq. (10) with respect to $\mathbf{f}^m$ is as follows

$$\begin{aligned}&\frac{\partial}{\partial\mathbf{f}^m}\left\{\sum_{m=1}^{M}\|\mathbf{\Omega}^m(\mathbf{f}^m-\mathbf{y})\|^2+\sum_{m=1}^{M}(\alpha^m)^2(\mathbf{f}^m)^\top\mathbf{L}^m\mathbf{f}^m\right.\\ &\left.+\mu\sum_{m=1}^{M}(\mathbf{f}^m)^\top\mathbf{\Omega}^m\sum_{p=1}^{M}\mathbf{\Omega}^p(\mathbf{f}^m-\mathbf{f}^p)\right\}\\ &=\mathbf{\Omega}^m(\mathbf{f}^m-\mathbf{y})+(\alpha^m)^2\mathbf{L}^m\mathbf{f}^m+\mu\mathbf{\Omega}^m\sum_{p=1}^{M}\mathbf{\Omega}^p(2\mathbf{f}^m-\mathbf{f}^p)=0.\end{aligned} \tag{11}$$

$$\Rightarrow\mathbf{f}^m=\left(\mathbf{\Omega}^m+(\alpha^m)^2\mathbf{L}^m+2\mu\mathbf{\Omega}^m\sum_{p=1}^{M}\mathbf{\Omega}^p\right)^{-1}\mathbf{\Omega}^m\left(\mathbf{y}+\mu\sum_{p=1}^{M}\mathbf{\Omega}^p\mathbf{f}^p\right). \tag{12}$$

In the second step, given fixed $\mathbf{F}$ and $\boldsymbol{\alpha}$, we can optimize $\{\mathbf{W}^m\}_{m=1}^M$. Then, the objective function in Eq. (9) can be re-written

as follows

$$\min_{\{\mathbf{W}^m\}_{m=1}^M} \sum_{m=1}^{M} (\alpha^m)^2 (\mathbf{f}^m)^\top \mathbf{L}^m \mathbf{f}^m + \lambda \sum_{m=1}^{M} \|\mathbf{W}^m\|_F^2, \quad \text{s.t.} \quad \sum_{j=1}^{N_e^m} W_{j,j}^m = 1, \quad \forall\, W_{j,j}^m \geq 0. \tag{13}$$

The partial derivative of Eq. (13) with respect to $\mathbf{W}^m$ is as follows

$$\frac{\partial}{\partial \mathbf{W}^m} \left\{ (\alpha^m)^2 (\mathbf{f}^m)^\top \mathbf{L}^m \mathbf{f}^m + \lambda \|\mathbf{W}^m\|_F^2 + \eta \left(\sum_{j=1}^{N_e^m} W_{j,j}^m - 1 \right) \right\} = 0. \tag{14}$$

$$\Rightarrow \quad \mathbf{W}^m = \frac{(\alpha^m)^2 \mathbf{\Lambda}^\top \mathbf{\Lambda} (\mathbf{D}_e^m)^{-1} - \eta \mathbf{I}^m}{2\lambda}, \qquad \eta = \frac{(\alpha^m)^2 \mathbf{\Lambda} (\mathbf{D}_e^m)^{-1} \mathbf{\Lambda}^\top - 2\lambda}{N_e^m}, \tag{15}$$

where $\mathbf{\Lambda} = (\mathbf{f}^m)^\top (\mathbf{D}_v^m)^{-\frac{1}{2}} \mathbf{H}^m$, and $\mathbf{I}^m \in \mathbb{R}^{N_e^m \times N_e^m}$ is an identity matrix.

In the third step, we optimize $\boldsymbol{\alpha}$ with fixed $\mathbf{F}$ and $\{\mathbf{W}^m\}_{m=1}^M$, and the problem in Eq. (9) can be re-written as follows

$$\min_{\boldsymbol{\alpha}} \sum_{m=1}^{M} (\alpha^m)^2 (\mathbf{f}^m)^\top \mathbf{L}^m \mathbf{f}^m, \quad \text{s.t.} \quad \sum_{m=1}^{M} \alpha^m = 1, \quad \forall\, \alpha^m \geq 0. \tag{16}$$

The partial derivative of Eq. (16) with respect to α^m is as follows

$$\frac{\partial}{\partial \alpha^m} \left\{ \sum_{m=1}^{M} (\alpha^m)^2 (\mathbf{f}^m)^\top \mathbf{L}^m \mathbf{f}^m + \tau \left(\sum_{m=1}^{M} \alpha^m - 1 \right) \right\} = 0. \tag{17}$$

$$\Rightarrow \quad \alpha^m = -\frac{\tau}{(\mathbf{f}^m)^\top \mathbf{L}^m \mathbf{f}^m}, \qquad \tau = -\frac{2 \prod_{m=1}^{M} (\mathbf{f}^m)^\top \mathbf{L}^m \mathbf{f}^m}{\sum_{m=1}^{M} \prod_{p=1, p\neq m}^{M} (\mathbf{f}^p)^\top \mathbf{L}^p \mathbf{f}^p}. \tag{18}$$

The alternating optimization process is repeated until convergence. The entire process of the above-mentioned method is summarized in Algorithm 1. In Fig. 4, we plot the change of the objective function values of Eq. (9) using different iteration numbers and the learned weights for hyperedges in the AD vs. NC classification, with $\mu = 10$ and $\lambda = 10$ for illustration. From Fig. 4 (top), it can be seen that the objective function value decreases rapidly within 5 iterations, illustrating the fast convergence of the proposed optimization algorithm. Fig. 4 (bottom) shows that the learned weights for different hyperedges vary significantly, implying that many hyperedges could be less discriminative in reflecting the true structure of data. In such a case, learning the optimal weights from data, as we do in this study via Eq. (9), provides an efficient way to suppress the contribution of hyperedges that are less important.

2.5. Multi-view label fusion

For a new testing sample $\mathbf{z}$, we now compute the weighted mean of its class probability scores $\{f_{\mathbf{z}}^m\}_{m=1}^M$ for making a final classification decision. Specifically, its class label can be obtained via

$$l(\mathbf{z}) = \text{sign}\left(\sum_{m=1}^{M} \frac{\alpha^m \times f_{\mathbf{z}}^m}{\gamma} \right), \tag{19}$$

Algorithm 1: View-aligned hypergraph classification.

Input: Labeled data with MRI, PET and CSF modalities, class label vector $\mathbf{y}$, and parameters μ and λ.

Step 1: Initialization

1.1: Group data into several views according to modalities, and construct the index matrix $\mathbf{\Omega}^m (m = 1, \ldots, M)$;

1.2: Construct multiple sets of hyperedges based on sparse representation coefficients in each of M view spaces, and compute corresponding matrices $\mathbf{H}^m$, $\mathbf{D}_v^m$ and $\mathbf{D}_e^m$;

1.3: Set $\mathbf{W}^m (m = 1, \ldots, M)$ as a diagonal matrix and $\boldsymbol{\alpha}$ with initial values;

repeat

  Step 2: Label update. Compute $\mathbf{F} = [\mathbf{f}^1, \ldots, \mathbf{f}^M]$ using Eq. (12);

  Step 3: Hyperedge weight update. Update the hyperedge weight $\mathbf{W}^m (m = 1, \ldots, M)$ based on Eq. (15);

  Step 4: View weight update. Compute the view weight $\boldsymbol{\alpha}$ via Eq. (18);

until *convergence*;

Output: $\mathbf{F}$, $\{\mathbf{W}^m\}_{m=1}^M$, and $\boldsymbol{\alpha}$.

where $\gamma = \sum_{m=1}^{M} \alpha^m$, and α^m is the optimal weight of the mth view learned from VAHC defined in Eq. (9).

It is worth noting that if $\mathbf{z}$ has missing values in a specific modality, the weights for related views associated with this modality will be 0. For instance, the weights for the views of "PET+MRI+CSF", "MRI+CSF" and "CSF" will be zeros if there are missing CSF data in the testing sample $\mathbf{z}$.

3. Results

Here, we present the competing methods (Section 3.1) and experimental settings (Section 3.2), followed by experimental results of our method in comparison to baseline methods (Section 3.3) and several state-of-the-art methods (Section 3.4). We further compare the computational costs of different methods in Section 3.5.

3.1. Methods for comparison

We first compare the proposed VAHL method with four baseline approaches based on data imputation techniques, including 1) Zero (missing values filled with zeros), 2) k-Nearest Neighbor (KNN) (Hastie et al., 1999; Troyanskaya et al., 2001; Hastie et al., 2005), 3) Expectation Maximization (EM) (Schneider, 2001), and 4) Singular Value Decomposition (SVD) (Golub and Reinsch, 1970). Assuming that the feature values are collected in the form of a matrix (as shown in Fig. 1), four baseline imputation-based methods are briefly summarized below.

1) In Zero method, missing values are filled with zeros. If data are normalized to have a mean of zero and unit standard deviation, this method is equivalent to the mean-value imputation method. That is, the missing feature values are filled with the means of corresponding feature values available in the same row.
2) In KNN method (Hastie et al., 1999; Troyanskaya et al., 2001), each missing value is filled with the weighted mean of its k-nearest neighbor columns. Specifically, we first adopt KNN to identify the feature columns that are most similar to the one with missing values. Those missing values are then filled in with the weighted mean of the values in the neighbor columns. Following (Thung et al., 2014), the weight for a specific neighbor column is inversely proportional to the Euclidean distance

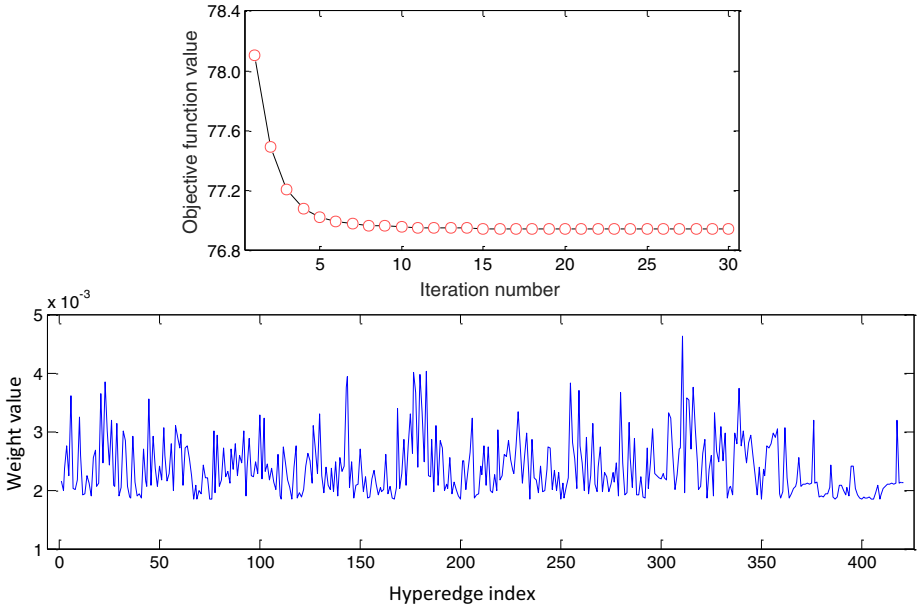

Fig. 4. Objective function values with respect to different iterations (top), and the learned weights for hyperedges (bottom) in AD vs. NC classification, with $\mu = 10$ and $\lambda = 10$.

between the neighbor column and the column with missing values.

3) In EM method (Schneider, 2001), missing values are imputed using the EM algorithm. Specifically, in the E step, we estimate the mean and the covariance matrix from the feature matrix, with missing values filled with the estimates from the previous E step (or initialized as zeros). In the M step, we assign the conditional expectation values to the missing elements based on the available values, the estimated mean, and the covariance. Next, we re-estimate the mean and the covariance according to the filled feature matrix. These two steps are repeated until convergence.
4) In SVD method (Golub and Reinsch, 1970), missing values are iteratively filled-in using the matrix completion technique with low-rank approximation. That is, some initial guesses (*e.g.*, zeros) are first assigned to the missing values and the method of SVD is then adopted to obtain a low-rank approximation of a filled-in matrix. Next, we update the missing elements with their corresponding values in the low-rank estimation matrix. Then, we perform SVD to the updated matrix again, and such processes are repeated until convergence.

The proposed VAHL method is further compared with six state-of-the-art methods: 1) two Ensemble based methods (Ingalhalikar et al., 2012) using weighted average (denoted as Ensemble-1) and average (denoted as Ensemble-2) strategies, respectively; 2) two incomplete multi-source feature (iMSF) learning methods (Yuan et al., 2012) with square loss (denoted as iMSF-1) and logistic loss (denoted as iMSF-2); 3) an incomplete source-feature selection (iSFS) method (Xiang et al., 2014); and 4) a matrix shrinkage and completion (MSC) method (Thung et al., 2014).

1) In the Ensemble based method (Ingalhalikar et al., 2012), an ensemble classification technique is adopted to fuse multiple classifiers by using different subsets of samples with complete data. Specifically, this method first divides the data into different subsets, and then selects relevant features using signal-to-noise ratio coefficient filter algorithm (Guyon and Elisseeff, 2003). Based on the selected features, a linear discriminant analysis (LDA) (Scholkopft and Mullert, 1999) classifier is constructed for each subset, followed by the fusion of classification results of multiple LDA classifiers to make a final decision for a testing subject. According to different fusion strategies, there are two versions of this method. The first one, denoted as Ensemble-1, is based on weighted averaging, where each classifier is assigned a specific weight based on its classification error on the training data. In the second approach (*i.e.*, Ensemble-2), all classifiers are assigned equal weights.
2) The iMSF method (Yuan et al., 2012) is a multi-view based method. Similar to our data grouping technique, iMSF first partitions subjects into several views, and a specific classifier is constructed in each view. A structural sparse learning model is then developed to select a common set of features among these tasks. Finally, an ensemble model is used to combine all models together. There are two versions of iMSF based on different loss functions, *i.e.*, the least square loss (denoted as iMSF-1) and the logistic loss (denoted as iMSF-2).
3) As another multi-view based method, iSFS (Xiang et al., 2014) first partitions subjects into several views according to the availability of data modalities. A bi-level (*i.e.*, both feature-level and view-level) feature learning model is proposed to learn the optimal weights for both features and views.
4) The MSC method (Thung et al., 2014) is a matrix completion based method. In MSC, the feature and the target output matrices are first combined into a large matrix that are partitioned into smaller sub-matrices, and each sub-matrix consists of samples with complete features (corresponding to a certain combination of modalities) and target outputs. A multi-task sparse learning method is applied to select informative features and samples, resulting in a shrunk version of the original matrix. The missing features and unknown target outputs of the shrunk matrix is then completed simultaneously, by using an EM imputation method (Schneider, 2001) or a fixed-point continuation method (Ma et al., 2011).

3.2. Experimental settings

In the experiments, we perform four classification tasks, including AD vs. NC, pMCI vs. NC, MCI vs. NC, and pMCI vs. sMCI classification. A 10-fold cross-validation strategy is used for performance evaluation. Specifically, all subjects are partitioned into

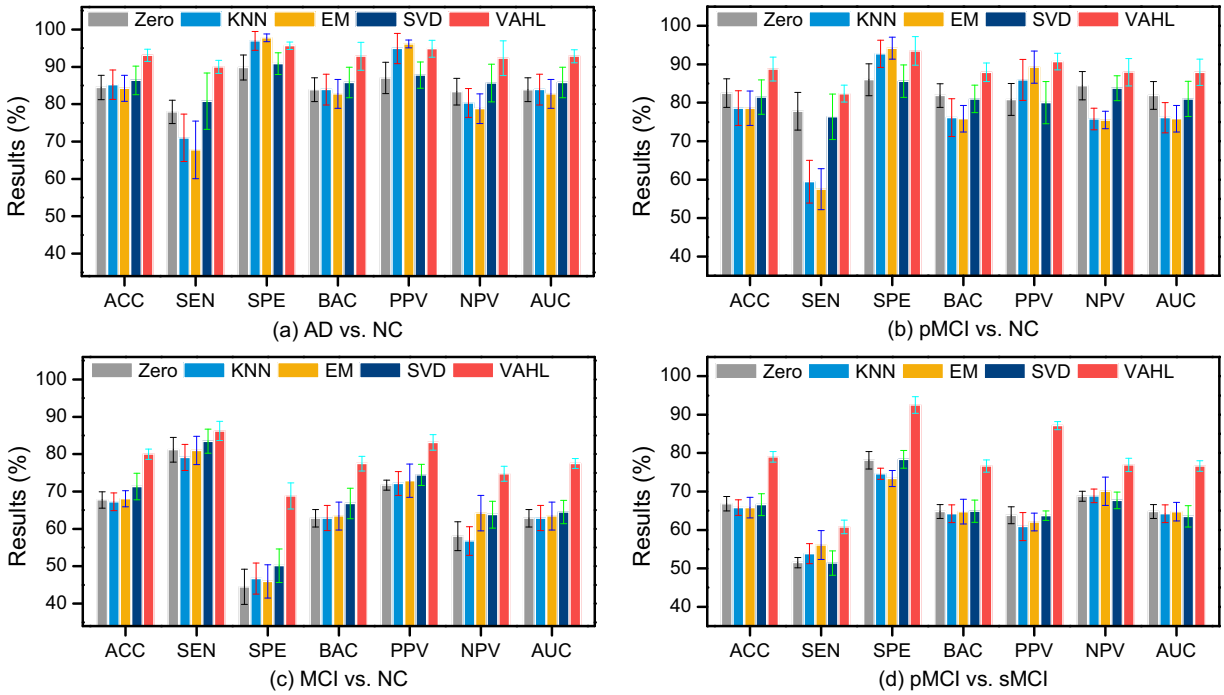

Fig. 5. Comparison between the proposed VAHL method and four baseline methods in four classification tasks.

10 subsets with roughly equal size. Each time one subset is designated as the testing data and the rest subsets as the training data. This process is repeated 10 times to avoid any bias introduced by random partitioning of the data, and finally the mean classification results are reported.

We adopt seven metrics for performance evaluation, including the classification accuracy (ACC), sensitivity (SEN), specificity (SPE), balanced accuracy (BAC), positive predictive value (PPV), negative predictive value (NPV) and the area under the receiver operating characteristic curve (AUC) (Fletcher et al., 2012). Denote TP, TN, FP and FN as true positive, true negative, false positive and false negative, respectively. These evaluation metrics are defined as: ACC=(TP+TN)/(TP+TN+FP+FN), SEN=TP/(TP+FN), SPE=TN/(TN+FP), BAC=(SEN+SPE)/2, PPV=TP/(TP+FP), and NPV=TN/(TN+FN).

To optimize the parameters for different methods, we further perform an inner 10-fold cross-validation using the training data. That is, each training subset is further divided into 10 subsets for cross-validation parameter selection (Xiang et al., 2014). The parameters in Eq. (9) (*i.e.*, μ and λ) are chosen from $\{10^{-3}, 10^{-2}, \cdots, 10^{4}\}$, while the iteration number in the proposed alternating optimization algorithm for Eq. (9) is empirically set to 20. Multiple parameter values for β in Eq. (4) are set to $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}]$ for constructing multiple sets of hyperedges for each hypergraph (*w.r.t.* each view) in VAHL. The parameter k for KNN is chosen from {3, 5, 7, 9, 11, 15, 20}, the rank parameter is chosen from {5, 10, 15, 20, 25, 30} for SVD, and the parameter λ for iMSF is chosen from $\{10^{-5}, 10^{-4}, \cdots, 10^{1}\}$.

3.3. Comparison with baseline methods

We first compare VAHL with imputation methods, including Zero, KNN (Hastie et al., 1999; Troyanskaya et al., 2001), EM (Schneider, 2001) and SVD (Golub and Reinsch, 1970). In Fig. 5, we report mean results as well as standard deviations achieved by different methods in four classification tasks, *i.e.*, AD vs. NC, pMCI vs. NC, MCI vs. NC, and pMCI vs. sMCI classification. From Fig. 5, we can observe that VAHL consistently outperforms those four baseline methods in terms of seven evaluation criteria.

3.4. Comparison with state-of-the-art methods

We further compare VAHL with several state-of-the-art methods, including Ensemble-1 and Ensemble-2 (Ingalhalikar et al., 2012), iMSF-1 and iMSF-2 (Yuan et al., 2012), iSFS (Xiang et al., 2014), and MSC (Thung et al., 2014). It is worth noting that iSFS first selects informative features from the original feature space, and then utilizes Random Forest classifier for classification. MSC utilizes matrix completion technique to simultaneously impute those missing values and unknown target outputs. The results for four classification tasks are reported in Table 2 and Table 3, where the best results are marked in boldface. In these tables, results of iSFS (Xiang et al., 2014) and MSC (Thung et al., 2014) are directly taken from their respective papers. From these two tables, we can observe that, in AD vs. NC, pMCI vs. NC, MCI vs. NC and pMCI vs. sMCI classification, our VAHL method generally outperforms the other methods in terms of ACC, SEN, SPE and AUC. For instance, in AD vs. NC classification, VAHL achieves a 4.6% improvement in terms of ACC compared with other methods. It is worth noting that both iSFS and VAHL learn the optimal weights for different views from data. Table 2 shows that, compared with iSFS, VAHL achieves much better results in AD vs. NC classification, and comparable results in pMCI vs. NC classification. The improvements given by VAHL can be attributed to the capability in modeling coherence among different views.

We further use the McNemars test (Dietterich, 1998) to assess whether the difference in performance between our proposed method and each competing method is significant, with the corresponding p-values reported in Table 4. These results show that our proposed method performs significantly better than the compared methods, as demonstrated by very small p-values (< 0.001).

3.5. Computational costs

Fig. 6 lists the computational costs of different methods in AD vs. NC classification. As shown in Fig. 6, the computational cost of VAHL is less than that of iMSF-2, and is comparable to SVD and iMSF-1. Compared with Zero, KNN, and EM methods, VAHL needs

Table 2
Comparison with the state-of-the-art methods in AD vs. NC and pMCI vs. NC classification.

Method	AD vs. NC				pMCI vs. NC			
	ACC (%)	SEN (%)	SPE (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
Ensemble-1	83.03	78.54	86.72	89.82	73.92	71.61	75.58	78.88
Ensemble-2	81.07	76.37	84.94	87.39	71.14	68.08	73.33	74.98
iMSF-1	86.41	76.91	94.24	85.57	82.53	69.32	92.11	80.71
iMSF-2	86.97	75.78	93.90	86.34	83.29	71.37	92.11	81.74
iSFS	88.48	88.95	88.16	88.56	89.86	**99.15**	84.00	91.57
MSC	88.50	83.70	92.70	94.40	–	–	–	–
VAHL	**93.10**	**90.00**	**95.65**	**94.83**	**89.95**	89.35	**93.48**	**92.00**

Table 3
Comparison with the state-of-the-art methods in MCI vs. NC and pMCI vs. sMCI classification.

Method	MCI vs. NC				pMCI vs. sMCI			
	ACC (%)	SEN (%)	SPE (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
Ensemble-1	62.58	65.42	57.73	64.40	68.10	55.44	77.77	64.60
Ensemble-2	61.61	64.16	57.28	62.07	65.56	51.15	75.41	61.78
iMSF-1	70.64	81.62	54.42	63.02	65.82	56.90	72.38	68.20
iMSF-2	71.61	82.83	54.73	63.78	64.55	56.85	70.22	66.00
MSC	71.50	75.30	64.90	77.30	–	–	–	–
VAHL	**80.00**	**86.19**	**68.78**	**80.49**	**79.00**	**60.80**	**92.53**	**79.66**

Table 4
The *p*-values in the McNemars test between the performances of the proposed method and each competing method in four classification tasks

Method	AD vs. NC	pMCI vs. NC	MCI vs. NC	pMCI vs. sMCI
Zero	0.0016	0.0021	0.0019	0.0030
KNN	0.0014	0.0026	0.0013	0.0033
EM	0.0028	0.0024	0.0017	0.0031
SVD	0.0027	0.0020	0.0024	0.0035
Ensemble-1	0.0038	0.0039	0.0011	0.0028
Ensemble-2	0.0032	0.0035	0.0013	0.0022
iMSF-1	0.0040	0.0042	0.0022	0.0023
iMSF-2	0.0039	0.0043	0.0020	0.0019

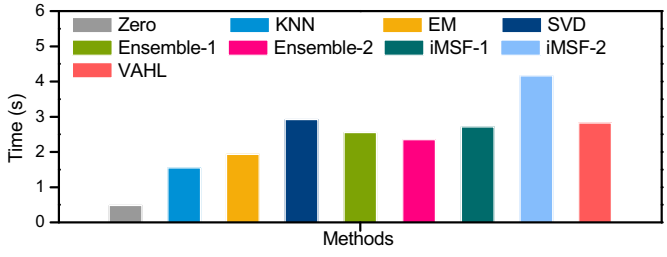

Fig. 6. Run time comparison between the proposed VAHL method and competing methods in AD vs. NC classification.

more computational time, due to the time spent on the construction of multiple hypergraphs. Overall, the computational cost of our method is reasonable and acceptable in practical applications.

4. Discussion

We first investigate the optimal weights for different views learned from our proposed VAHC model in Section 4.1, and then evaluate the influence of two regularization parameters in Eq. (9) in Section 4.2. In Section 4.3, we study the influence of different similarity measurement for hyperedge construction, including the proposed sparse representation and conventional Euclidean distance based measurements. We also study the influence of the proposed view-centralized regularizer on the learning performance

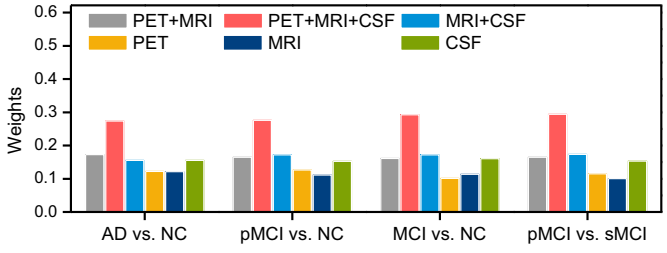

Fig. 7. Optimal weights of different views (*i.e.*, "PET+MRI", "PET+MRI+CSF", "MRI+CSF", "PET", "MRI", and "CSF") learned from the proposed view-aligned hypergraph classification model in four classification tasks.

in Section 4.4. In Section 4.5, we further show the results using complete data in the ADNI-1 database.

4.1. Learned weights for different views

Now we show the optimal weights for different views learned from the proposed VAHC model defined in Eq. (9), with results given in Fig. 7. From Fig. 7, we can observe that the weights for the view of "PET+MRI+CSF" are much larger than those of the other five views in four classification tasks. This indicates that the view that contains the combination of MRI, PET, and CSF data can provide more discriminative information, compared with the other views. Among three views that contain only one single modality data, Fig. 7 indicates that the weights for the view of "CSF" are generally larger than those for the views of "MRI" and "PET". This implies that CSF could be comparatively more effective biomarkers in distinguishing AD/MCI patients from the whole population, compared with MRI and PET data.

4.2. Influence of regularization parameters

In the proposed classification model in Eq. (9), there are two parameters (*i.e.*, μ and λ) for our proposed view-aligned regularizer and the hyperedge weight regularizer, respectively. We have evaluated the influences of those two regularization parameters on the performance of our method, with results shown in Fig. 8. The

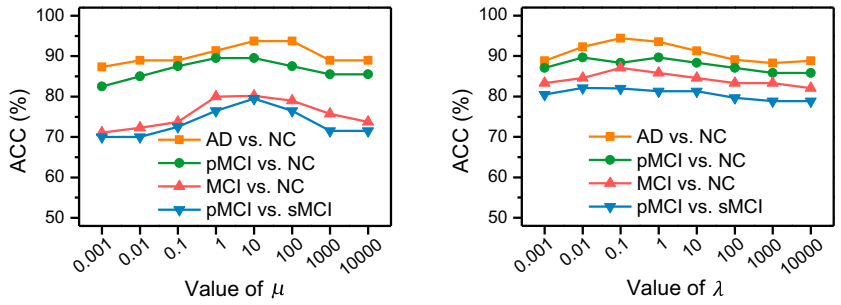

Fig. 8. Influence of the parameters (*i.e.*, μ and λ) on the proposed method.

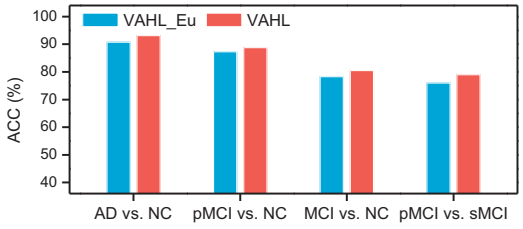

Fig. 9. Comparison between VAHL and VAHL_Eu. Here, VAHL and VAHL_Eu denote the proposed methods that adopt sparse representation coefficients and Euclidean distance as similarity measurements for constructing hyperedges, respectively.

values of μ and λ are varied within $\{10^{-3}, 10^{-2}, \cdots, 10^{4}\}$. From Fig. 8, we can observe we can observe that, with different values of μ and λ, the classification accuracies fluctuate in a large range. For instance, as shown in Fig. 8 (left), VAHL generally achieves better results in terms of ACC when $1 \leq \mu \leq 10$. Also, Fig. 8 (right) indicates that the best results are usually obtained by VAHL using $0.1 \leq \lambda \leq 1$ in four classification tasks. These results imply that the proposed view-aligned regularizer and the hypergraph Laplacian regularizer play important roles in the VAHC model.

4.3. Sparse representation coefficients vs. Euclidean distance

We further compare VAHL (using a sparse representation based hypergraph construction approach) with the conventional method (denoted as VAHL_Eu) that uses the Euclidean distance as similarity measurement for constructing a hypergraph in each view space. For fair comparison, in VAHL_Eu, multiple neighbors are used for constructing hyperedges via the star expansion algorithm (Zien et al., 1999), where each centroid vertex is connected with its s-nearest neighbors. In the experiments, we adopt the neighbor size $s = [3, 5, 7, 9, 11, 15]$ for VAHL_Eu. Fig. 9 reports the classification accuracies achieved by VAHL and VAHL_Eu. From this figure, we can observe that VAHL consistently outperforms VAHL_Eu in four classification tasks. This demonstrates that, for hypergraph construction, the use of sparse representation brings performance improvement compared with that of the Euclidean distance. This can partly contribute to the global structure information conveyed by sparse representation coefficients (Wright et al., 2009).

4.4. Influence of the view-aligned regularizer

We also study the influence of the proposed view-aligned regularizer on the classification performance. We denote "VAHL_noVA" as the VAHL model without the view-aligned regularizer (*i.e.*, $\mu = 0$ in Eq. (9)), and perform experiments to compare VAHL and VAHL_noVA (with results shown in Fig. 10). It can be seen from Fig. 10 that VAHL outperforms VAHL_noVA in terms of accuracy in four classification tasks, implying that modeling the coherence among views via the proposed view-aligned regularizer can boost the classification performance of hypergraph based model.

4.5. Complete data vs. incomplete data

We further investigate whether methods using incomplete data can boost the learning performance, compared with those using only complete data (with PET, MRI, and CSF features). In the baseline ADNI-1 database, there are a total of 202 subjects that have complete data, including 51 AD, 42 pMCI, 57 sMCI, and 52 NC subjects. We compare the proposed VAHL method with both SVM and multi-kernel SVM (MHL_SVM) (Zhang et al., 2011) using complete data, with corresponding results shown in Fig. 11. Here, the concatenation of MRI, PET, and CSF features is used in SVM, while each of three data modalities is treated as a specific kernel in MHL_SVM. It is worth noting that our VAHL model has only one view (*i.e.*, "PET+MRI+CSF") in the case of using complete data. From Figs. 11 and 5, we can observe that the overall performance of methods using complete data is worse than that of the method using incomplete data, suggesting that utilizing more data can promote the AD/MCI diagnosis performance. Also, it can be seen from Fig. 11 that VAHL generally outperforms the conventional SVM and MHL_SVM, demonstrating that our method provides a better way to utilize multi-modality data for AD/MCI diagnosis.

5. Conclusions

In this paper, we propose a view-aligned hypergraph learning (VAHL) method using incomplete multi-modality data for AD/MCI diagnosis. Specifically, we first partition the original data into several views according to the availability of data modalities, and construct one hypergraph in each view using a sparse representation based hypergraph construction approach. We then develop a view-aligned hypergraph classification model to explicitly capture the underlying coherence among views, as well as automatically learn the optimal weights of different views from data. A multi-view label fusion method is employed to assemble the estimated class probability scores to arrive at a final classification decision. Results on the baseline ADNI-1 database (with MRI, PET, and CSF modalities) demonstrate the efficacy of our method in AD/MCI diagnosis. In this study, we employ all original features for hypergraph construction, while there may exist noisy or redundant information in original features. It is interesting to select those most informative features for subsequent hypergraph construction, which will be part of our future work. Also, we only perform experiments on the baseline ADNI-1 database with three data modalities. As a future work, we will evaluate the proposed method on more datasets, such as the ADNI-2 database and the dataset in the Computer-Aided Diagnosis of Dementia (CADDementia) challenge (Bron et al., 2015).

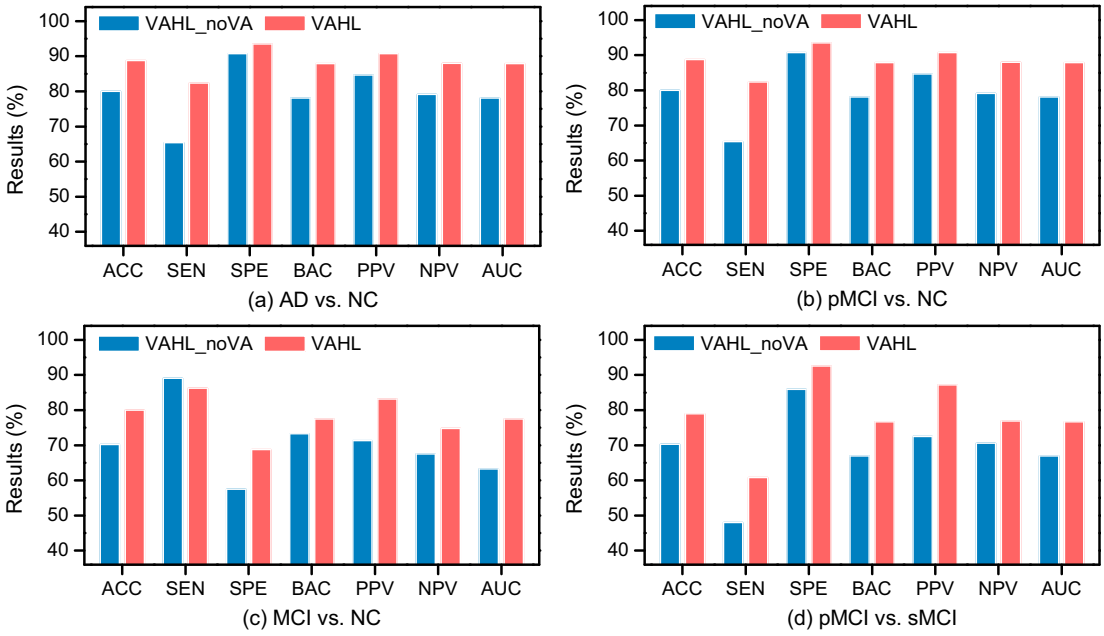

Fig. 10. Comparison between VAHL and VAHL_noVA. Here, VAHL_noVA denotes the proposed VAHL model without the view-aligned regularizer.

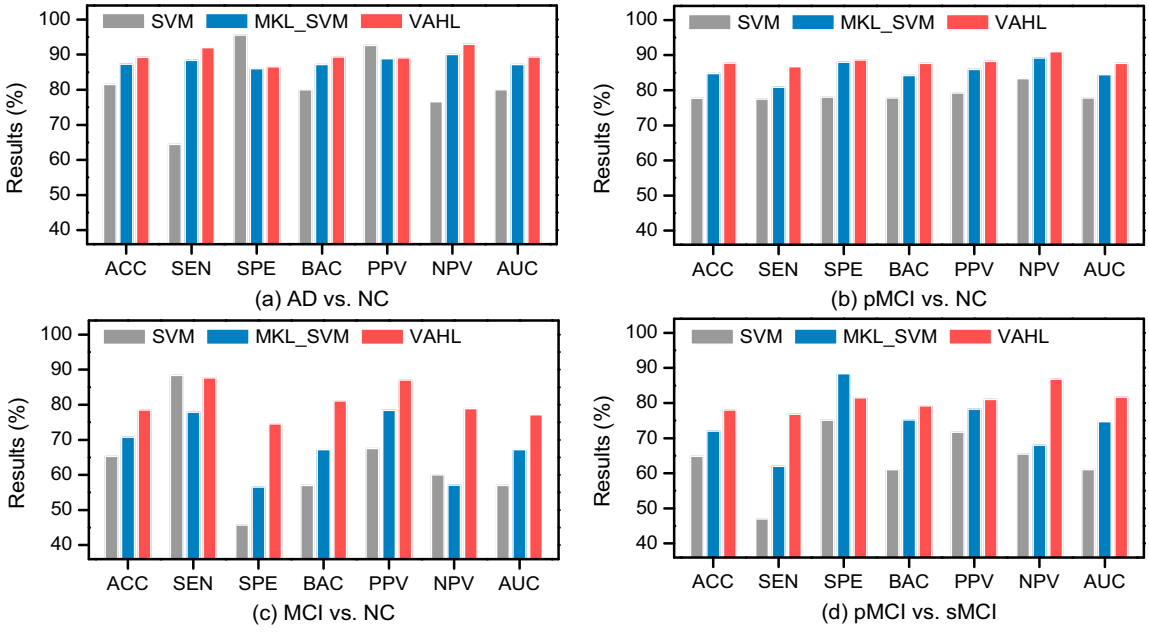

Fig. 11. Classification results achieved by different methods using complete data, where MHL_SVM denotes multi-kernel SVM.

Acknowledgments

This study was supported by NIH grants (EB006733, EB008374, EB009634, MH100217, AG041721, AG042599, AG010129, and AG030514).

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found online.[4]

[4] https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

References

Argyriou, A., Evgeniou, T., Pontil, M., 2008. Convex multi-task feature learning. Mach. Learn. 73 (3), 243–272.

Association, A., et al., 2013. 2013 Alzheimer's disease facts and figures. Alzheimer's Dementia 9 (2), 208–245.

Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., et al., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. Neuroimage 111, 562–579.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M., 2007. Forecasting the global burden of Alzheimer's disease. Alzheimer's Dementia 3 (3), 186–191.

Chetelat, G., Desgranges, B., De La Sayette, V., Viader, F., Eustache, F., Baron, J.-C., 2003. Mild cognitive impairment can FDG-PET predict who is to rapidly convert to Alzheimer's disease? Neurology 60 (8), 1374–1377.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56 (2), 766–781.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923.
Fletcher, R.H., Fletcher, S.W., Fletcher, G.S., 2012. Clinical Epidemiology: The Essentials. Lippincott Williams & Wilkins.
Foster, N.L., Heidebrink, J.L., Clark, C.M., Jagust, W.J., Arnold, S.E., Barbas, N.R., DeCarli, C.S., Turner, R.S., Koeppe, R.A., Higdon, R., 2007. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain 130 (10), 2616–2635.
Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q., 2012. 3-D object retrieval and recognition with hypergraph analysis. IEEE Trans. Image Process. 21 (9), 4290–4303.
Golub, G.H., Reinsch, C., 1970. Singular value decomposition and least squares solutions. Numerische Mathematik 14 (5), 403–420.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K., Minthon, L., 2006. Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. Lancet Neurol. 5 (3), 228–234.
Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. Mathematical Intell. 27 (2), 83–85.
Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., Botstein, D., 1999. Imputing Missing Data for Gene Expression Arrays.
Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frölich, L., Schönknecht, P., Ito, K., Mielke, R., Kalbe, E., 2002. Discrimination between alzheimer dementia and controls by automated analysis of multicenter FDG PET. Neuroimage 17 (1), 302–316.
Ingalhalikar, M., Parker, W.A., Bloy, L., Roberts, T.P., Verma, R., 2012. Using multiparametric data with missing features for learning patterns of pathology. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 468–475.
Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.
Kawarabayashi, T., Younkin, L.H., Saido, T.C., Shoji, M., Ashe, K.H., Younkin, S.G., 2001. Age-dependent changes in brain, CSF, and plasma amyloid β protein in the tg2576 transgenic mouse model of Alzheimer's disease. J. Neurosci. 21 (2), 372–381.
Liu, M., Zhang, D., Shen, D., 2016. Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans. Med. Imaging 35 (6), 1463–1474.
Ma, S., Goldfarb, D., Chen, L., 2011. Fixed point and bregman iterative methods for matrix rank minimization. Math Program 128 (1–2), 321–353.
Qiao, H., Zhang, H., Zheng, Y., Ponde, D.E., Shen, D., Gao, F., Bakken, A.B., Schmitz, A., Kung, H.F., Ferrari, V.A., et al., 2009. Embryonic stem cell grafting in normal and infarcted myocardium: serial assessment with MR imaging and PET dual detection. Radiology 250 (3), 821–829.
Qiao, L., Chen, S., Tan, X., 2010. Sparsity preserving projections with applications to face recognition. Pattern Recognit. 43 (1), 331–341.
Reiman, E.M., Langbaum, J.B., Tariot, P.N., 2010. Alzheimer'S prevention initiative: a proposal to evaluate presymptomatic treatments as quickly as possible. Biomark Med. 4 (1), 3–14.
Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J. Clim. 14 (5), 853–871.
Scholkopft, B., Mullert, K.-R., 1999. Fisher discriminant analysis with kernels. Neural Netw. Signal Process. IX 1, 1.
Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21 (11), 1421–1439.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.
Thung, K.-H., Wee, C.-Y., Yap, P.-T., Shen, D., 2014. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. Neuroimage 91, 386–400.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15 (1), 273–289.
Verma, R., Mori, S., Shen, D., Yarowsky, P., Zhang, J., Davatzikos, C., 2005. Spatiotemporal maturation patterns of murine brain quantified by diffusion tensor MRI and deformation-based morphometry. Proc. Natl. Acad. Sci. U.S.A. 102 (19), 6978–6983.
Wang, Y., Nie, J., Yap, P.-T., Shi, F., Guo, L., Shen, D., 2011. Robust deformable-surface-based skull-stripping for large-scale studies. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer, pp. 635–642.
Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J., 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS ONE 6 (10), e25446.
Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2), 210–227.
Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J., 2014. Bi-level multi-source learning for heterogeneous block-wise missing data. Neuroimage 102, 192–206.
Xue, Z., Shen, D., Karacali, B., Stern, J., Rottenberg, D., Davatzikos, C., 2006. Simulating deformations of MR brain images for validation of atlas-based segmentation and registration algorithms. Neuroimage 33 (3), 855–866.
Yang, J., Shen, D., Davatzikos, C., Verma, R., 2008. Diffusion tensor image registration using tensor geometry and orientation features. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 905–913.
Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J., 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. Neuroimage 61 (3), 622–632.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.
Zhang, J., Gao, Y., Gao, Y., Munsell, B., Shen, D., 2016. Detecting anatomical landmarks for fast Alzheimer's disease diagnosis. IEEE Trans. Med. Imaging.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20 (1), 45–57.
Zhou, D., Huang, J., Schölkopf, B., 2006. Learning with hypergraphs: clustering, classification, and embedding. In: Advances in Neural Information Processing Systems, pp. 1601–1608.
Zien, J.Y., Schlag, M.D., Chan, P.K., 1999. Multilevel spectral hypergraph partitioning with arbitrary vertex sizes. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 18 (9), 1389–1399.